Breast Cancer Biomarker Prediction Model Based on Principal Component Extraction and Deep Convolutional Network Integration Learning

Kun Ruan¹, Yuhao Peng¹, Yuhang Kang¹, Shun Zhao¹, Tanke Wang¹, Yuchi Zhang¹ and Tao Liu^{1,*}

¹ Sichuan Agricultural University, No. 46, Xinkang Road, Ya'an, Sichuan, China

Abstract. Effective extraction of characteristic information from sequencing data of cancer patients is an essential application for cancer research. Several prognostic classification models for breast cancer sequencing data have been established to assist patients in their treatment. However, these models still have problems such as poor robustness and low precision. Based on the convolutional network model in deep learning, we construct a new classifier PCA-1D LeNet-Ada (PLA) by using principal component extraction method, Le-Net convolution network, and Adaptive Boosting method. PLA predicts three biomarkers for breast cancer patients based on their somatic cell copy number variations and gene expression profiles.

1 Introduction

Determining the relationship between sequencing data and cancer characterization is significant for clinical decision-making and diagnosis and treatment. Breast cancer is the most common cause of death from cancer in women worldwide [1]. Based on progesterone receptor (PR), human epidermal growth factor receptor 2 (HER2), and estrogen receptor (ER), breast cancer can be divided into four molecular subtypes: Luminal A (ER+, PR+, and HER2-), Luminal B (ER+, PR±, and Her2±), Her2-overexpression (ER-, PR- and HER2+) and triple-negative (ER-, PR- and HER2-) [2, 3]. Among them, the prognostic effect of Luminal A and Luminal B subtypes is relatively good, while the prognostic effect of triple-negative and HER2-(HER2 enriched) tumors is inferior [4-6]. Accurately predicting the status of ER, PR and HER2 can provide excellent help for breast cancer treatment and prognosis selection.

In recent years, many scholars have combined deep learning with gene sequencing data for cancer analysis experiments. For example, Xiaofan Ding et al. tested the possible correlation between recurrent CNV in the genome and the huge risk of cancer [7], and Md. Mohaiminul Islam et al. established a prediction of breast cancer molecular subtypes based on changes in somatic copy number Deep learning model [8].

At present, the main factors affecting the accuracy of cancer prognosis models are the high heterogeneity of biological data and the difficulty of achieving stable extraction based on existing data. On the problem of the 1D signal, convolutional neural network has been proved to be effective. In 2015, Kiranyaz proposed an adaptive 1D convolutional neural network (1D CNN) for patient-specific ECG data, which promoted the application of 1D convolutional networks in the medical field [9]. In

terms of breast cancer prediction, there is still a lack of network models that integrate multiple omics data to predict the receptor status of different patients accurately. Based on this, this article proposes a PLA model to explore the relationship between somatic cell copy number variation, gene expression, and receptor status in different individuals. We established the prediction model of breast cancer ER, PR, HER2 status.

2 MATERIALS AND METHODS

2.1 Data set and data preprocessing

This article selects Breast Cancer (METABRIC, Nature 2012 & Nat Commun 2016) as the experimental data set. The information of this data set is measured by Illumina Human v3 microarray. It can be downloaded from cBioPortal, an analysis tool for exploring large cancer genome data sets.

Fig. 1. The proportion of breast cancer types in the data set

* Corresponding author: <mailto:tao666@qq.com>

The data set contains 2509 patient samples, and the gene expression of each sample is composed of genome profiles (somatic mutation [targeted sequencing], copy number variation, and gene expression) and corresponding clinical data.

Figure 1 shows the proportion of breast cancer types in the data set. In this data set, invasive ductal carcinoma of the breast accounted for 74.3% of the total sample number, mixed ductal and lobular breast carcinoma accounted for 10.7% of the total sample number, invasive lobular carcinoma of the breast accounted for 7.7% of the total sample number, and other subtypes of breast cancer and Unlabeled breast cancer cases accounted for 7.3% of the total sample number.

Before constructing the predictive classification model, we performed a three-step preprocessing operation on the data set.

(1) Eliminate the separate components in the data set, and select Copy-number alterations from DNA copy (CNV) (22545 gene information), gene expression data (24367 gene information), and in clinical data, ER Status, PR Status, and HER2 Status are studied.

(2) Eliminate samples with NaN values in the current data set

(3) Perform Z-Score standardization on the original CNV data and RNA_Seq data to eliminate the difference between dimensions.

After data preprocessing operations, we got 1885 available samples.

2.2 Data feature extraction method

2.2.1 Principal component analysis

Principal component analysis (PCA) is a multivariate technique that can significantly reduce the number of variables while still retaining most of the original information. PCA is often used for data dimensionality reduction, lossy data compression, feature extraction, and data visualization [10].

2.2.2. Deep Learning Model

Fig. 2. Le-Net Model Framework under Marker Classification

Compared with previous shallow machine learning models, convolutional neural networks have unprecedented advantages in feature extraction, expression, and model fitting. For more complex tasks, such as segmentation of pathological images, deep and complex neural networks are often used. For numerical data tasks, calculation speed is also a performance index

that must be considered while pursuing prediction accuracy. Combining the above reasons, we chose the LeNet-5 network to extract data features, which can improve the calculation speed while predicting with high precision. In this paper, the constructed model for predicting clinical outcomes of breast cancer patients is named PCA-1DLenet. Among them, the dimensionality reduction feature data obtained in the previous step is used as the input of the classification network.

The hierarchical structure of LeNet-5 includes three types: convolutional layer, pooling layer, and fully connected layer. Among them, the convolutional layer and the pooling layer perform feature extraction. The fully connected layer maps the extracted features to the final output to achieve the final classification.

To limit the over-fitting of the network, we adopt the random Dropout method. Randomly reset part of the weight of the hidden layer to zero to adjust the dependency between nodes. At the same time, we use the ReLU function as the activation function and the cross-entropy function as the loss function. See Table 1 for the parameter settings of the PLA model.

Table. 1. PLA model parameter table

Layer	Input size	kernel size	Sampling area	kernel type
Conv1	1×1059	5×5	-	10
Pool1	1×1055×10	-	2×2	-
Conv2	1×528×10	3×3	-	16
Pool2	1×526×16	-	2×2	-
Fc1	4208	-	-	-
Fc2	256	-	-	-
Output	128	-	-	-
Layer	Sample type	Neurons Number	Sliding step	Output size
Conv1	-	1×1055	1×1	1×1055×10
Pool1	10	1×528	2×2	1×528×10
Conv2	16	1×526	1×1	1×526×16
Pool2	16	1×263	2×2	1×263×16
Fc1	-	256	-	256
Fc2	-	128	-	128
Output	-	2	-	2
Layer	Trainable parameter	Connections Number	Dropout	
Conv1	(5×5+1)×10	(5×5+1)×10×1×1055	-	
Pool1	2×10	(2×2+1)×10×1×528	-	
Conv2	(3×3+1)×16	(3×3+1)×16×1×526	-	
Pool2	2×16	(2×2+1)×16×1×263	-	
Fc1	256+1	(256+1)×4208	0.5	
Fc2	128+1	(128+1)×256	0.5	
Output	2+1	(2+1)×128	-	

2.2.3 AdaBoost method

Fig. 3. AdaBoost operation steps

Adaptive Boosting is a machine learning method proposed by Yoav Freund and Robert Schapire [11]. The AdaBoost method is sensitive to noisy data and abnormal data. It is essentially an iterative algorithm, adding a new weak classifier in each round until it reaches a sufficiently small error rate. For each training, if a sample point has been accurately classified, then in the construction of the next training set, its probability of being selected is reduced; on the contrary, if a sample point is not accurately classified, then its weight is obtained improve. In this way, the AdaBoost method can focus on samples that are more difficult to distinguish (more informative).

In this paper, a new 1DLeNet weak classifier is added in each iteration of AdaBoost, and the first

prediction result is corrected by calculating the training error and updating the original weight. According to multiple experimental tests, in the case of 10 iterations, the overall prediction effect of the PCA model on the test set data reaches the best level.

2.3 Biomarker prognostic model based on principal component extraction and deep convolutional network-integrated learning

This research aims to combine principal component feature extraction methods with deep learning techniques to learn more characteristic features from sequencing data such as copy number variation and gene expression and construct a more useful classification of breast cancer clinical outcome predictions.

Based on PCA, LeNet-5 network, and Adaptive Boosting method, we built a PLA classification model. The overall structure of the PLA is shown in Figure 4. First, the PCA dimensionality reduction method is used to reduce the dimensionality of the characteristics of copy number variation and gene expression data. Then, use 1D LeNet deep convolutional network technology to build a classification system PCA-1DLeNet.

Fig. 4. PLA model implementation framework

Finally, this study uses 10 PCA-1DLeNets as weak classifiers and uses Adaptive Boosting to build an ensemble classifier. Through ensemble learning, it learns the negative and positive classification features of three receptors of breast cancer, from dimensionality reduction features.

3 RESULT

3.1 Evaluation indicators

Due to the high imbalance in the distribution of the number of patients with negative and positive receptors

in the original data set, for example, out of 1885 samples, 1653 samples were HER2 negative, and 232 samples were HER2 positive. The Matthews correlation coefficient is widely used in this situation [1]. Therefore, this study uses accuracy (ACC), Matthews correlation coefficient (MCC), and the verification deviation correction area (AUC) under the ROC curve as the main performance evaluation indicators for evaluating the diagnosis. The calculation formula of MCC is as follows:

$$MCC = \frac{t^+ t^- - f^+ f^-}{\sqrt{(t^+ + f^+)(t^+ + f^-)(t^- + f^+)(t^- + f^-)}}$$

Among them, true positive is a true negative, false positive is a false negative, and the formulas for

accuracy rate ACC, specificity SP, and sensitivity SN are as follows:

$$ACC = \frac{t^+ + t^-}{t^+ + t^- + f^+ + f^-} \quad SP = \frac{t^-}{t^- + f^+} \quad SN = \frac{t^+}{t^+ + f^-}$$

To avoid sampling bias and ensure the accuracy of the conclusion, this paper uses 10-Fold cross-validation to compare and analyze the results of breast cancer receptor classification.

3.2 Comparison of experimental results

In this paper, 300 iterations of training are performed on the data set. During the ER training process, as the average cross-entropy loss of the training set continues to decrease, its accuracy is continuously improved. At the same time, the average cross-entropy loss of the test set keeps the same trend as the accuracy rate and the training set. When it occurs after 100 iterations, the accuracy of the training set gradually converges, and the accuracy of the test set stabilizes at about 92%.

HER2 and PR also showed better training conditions. After HER2 iterated to 100 times, the accuracy of the training set gradually converged to 1, and the accuracy of the test set stabilized at about 90%. After the PR iteration reaches 120 times, the accuracy of the training set gradually converges, and the accuracy of the test set stabilizes at about 75%.

The experimental results prove that the classifier we built has a good effect on the training set and test set for the classification of ER, HER2, and PR. Among them, the ACC scores of ER and HER2 are above 0.9, and the ACC of PR is above 0.7. And the specific SN sensitivity SP of the model established for the three receptors is also maintained at a high level.

Table 2. PLA forecast results

	ER	HER2	PR
ACC	0.923	0.900	0.730
MCC	0.783	0.452	0.456
AUC	0.884	0.727	0.681
SP	0.958	0.392	0.766
SN	0.810	0.970	0.688

To verify the effectiveness of the PLA model, we compared the PLA model with four other mature classification models (SVM, Dtree, KNN, Perceptron) [12-15] for the classification of the three receptors: ACC, MCC, and AUC.

Fig. 5. Comparison of ACC and MCC of five models

It can be found that the accuracy of the PLA model in the ER and HER2 receptors is significantly higher than the other four different algorithms, which shows the excellent classification performance of PLA. In terms of MCC scores, compared to the other four algorithms, the PLA model showed better predictability in three different receptors. Among them, the highest value in ER reaches 0.78, the value of HER2 reaches 0.45, and the value of PR reaches 0.46. This also confirms the robustness of the PLA model's classification results under unbalanced positive and negative samples.

4 DISCUSSION

This paper compares the classification results with traditional models, and the results show that the PLA model is better for molecular receptor classification. There are two main reasons for this. First, the amount of data in the data set we use is more sufficient than previous studies; second, the application of convolutional networks to the field of clinical, histological image recognition has shown better accuracy. We extend this model to In terms of data analysis, according to the evaluation test results, we have achieved excellent results in the classification of breast cancer molecular receptors.

However, due to the complex structure of the deep learning model, it is still difficult to explore the features that specifically affect the classification results. In the future, we will consider using a newer and deeper network as Backbone to experiment with the classification effect.

5 CONCLUSION

The purpose of this study is to use breast cancer gene expression and copy number variation data to explore the relationship between different molecular receptors. Based on deep learning technology, we have established the PLA breast cancer classification model and realized the prediction of the molecular markers ER, PR, and Her-2 of breast cancer. The classification accuracy rate of the model on the test set is reasonable, stable at about 92%; among them, the positive prediction of ER, PR is good, the classification accuracy of tumors with negative ER, PR, and Her-2 or Her-2 is average; Besides, the classification accuracy of ER, HER2, and PR receptors were 0.92, 0.90, and 0.73, respectively.

The main work of the article is on the construction of the overall prediction model. In the follow-up, we will further integrate the classification results of ER, PR, and HER2 receptors to achieve high-precision prediction of molecular subtypes. Our PLA model can be extended to the patient data of various common cancers. The establishment of this type of model will play a key role in cancer treatment and prognostic selection.

References

1. T. J. Key, P. K. Verkasalo, and E. Banks, "Epidemiology of breast cancer," *The lancet oncology*, vol. 2, no. 3, pp. 133-140, 2001.
2. L. A. Carey *et al.*, "Race, breast cancer subtypes, and survival in the Carolina Breast Cancer Study," *Jama*, vol. 295, no. 21, pp. 2492-2502, 2006.
3. K. M. O'Brien *et al.*, "Intrinsic breast tumor subtypes, race, and long-term survival in the Carolina Breast Cancer Study," *Clinical Cancer Research*, vol. 16, no. 24, pp. 6100-6110, 2010.
4. K. R. Bauer, M. Brown, R. D. Cress, C. A. Parise, and V. Caggiano, "Descriptive analysis of estrogen receptor (ER)-negative, progesterone receptor (PR)-negative, and HER2-negative invasive breast cancer, the so-called triple-negative phenotype: a population-based study from the California cancer Registry," *cancer*, vol. 109, no. 9, pp. 1721-1728, 2007.
5. R. Haque *et al.*, "Impact of breast cancer subtypes and treatment on survival: an analysis spanning two decades," *Cancer Epidemiology and Prevention Biomarkers*, vol. 21, no. 10, pp. 1848-1855, 2012.
6. S. A. Joslyn, D. H. Gesme Jr, and C. F. Lynch, "Estrogen and progesterone receptors in primary breast cancer," *The Breast Journal*, vol. 2, no. 3, pp. 187-196, 1996.
7. X. Ding, S.-Y. Tsang, S.-K. Ng, and H. Xue, "Application of machine learning to development of copy number variation-based prediction of cancer risk," *Genomics insights*, vol. 7, p. GEI. S15002, 2014.
8. M. M. Islam, R. Ajwad, C. Chi, M. Domaratzki, Y. Wang, and P. Hu, "Somatic copy number alteration-based prediction of molecular subtypes of breast cancer using deep learning model," in *Canadian Conference on Artificial Intelligence*, 2017: Springer, pp. 57-63.
9. S. Kiranyaz, T. Ince, R. Hamila, and M. Gabbouj, "Convolutional neural networks for patient-specific ECG classification," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2015: IEEE, pp. 2608-2611.
10. I. Jolliffe, *Principal component analysis*. Springer, 2011.
11. Y. Freund, R. Schapire, and N. Abe, "A short introduction to boosting," *Journal-Japanese Society For Artificial Intelligence*, vol. 14, no. 771-780, p. 1612, 1999.
12. M. G. Best *et al.*, "RNA-Seq of tumor-educated platelets enables blood-based pan-cancer, multiclass, and molecular pathway cancer diagnostics," *Cancer cell*, vol. 28, no. 5, pp. 666-676, 2015.
13. N. M. Crabtree, J. H. Moore, J. F. Bowyer, and N. I. George, "Multi-class computational evolution: development, benchmark evaluation and application to RNA-Seq biomarker discovery," *BioData mining*, vol. 10, no. 1, p. 13, 2017.
14. T. Nguyen, S. Nahavandi, D. Creighton, and A. Khosravi, "RNA-seq data analysis using nonparametric Gaussian process models," in *2016 International Joint Conference on Neural Networks (IJCNN)*, 2016: IEEE, pp. 5087-5093.
15. P. H. Williams, R. Eyles, and G. Weiller, "Plant microRNA prediction by supervised machine learning using C5. 0 decision trees," *Journal of nucleic acids*, vol. 2012, 2012.